Dual Adversarial Network for Unsupervised Ground/Satellite-to-Aerial Scene Adaptation

Jianzhe Lin
jianzhelin@ece.ubc.ca
University of British Columbia

Lichao Mou
lichao.mou@dlr.de
Technical University of Munich,
German Aerospace Center

Tianze Yu*
tian.yori90@gmail.com
University of British Columbia

Xiaoxiang Zhu
xiaoxiang.zhu@dlr.de
Technical University of Munich,
German Aerospace Center

Z. Jane Wang
zjanew@ece.ubc.ca
University of British Columbia

Figure 1: Examples of scenes from top-down views. From top to down are scenes from the satellite view, the aerial view, and the ground view. Scenes from the satellite view are with much lower resolution and clarity compared with the aerial view. Scenes from the ground view and the aerial view are with huge domain gap even with the consistent semantic labels.

ABSTRACT

Recent domain adaptation work tends to obtain a uniformed representation in an adversarial manner through joint learning of the domain discriminator and feature generator. However, this domain adversarial approach could render sub-optimal performances due to two potential reasons: First, it might fail to consider the task at hand when matching the distributions between the domains. Second, it generally treats the source and target domain data in the same way. In our opinion, the source domain data which serves the feature adaption purpose should be supplementary, whereas the target domain data mainly needs to consider the task-specific classifier [1]. Motivated by this, we propose a dual adversarial network for domain adaptation, where two adversarial learning processes are conducted iteratively, in correspondence with the feature adaptation and the classification task respectively. The efficacy of the proposed method is first demonstrated on Visual Domain Adaptation Challenge (VisDA) 2017 challenge, and then on two newly proposed Ground/Satellite-to-Aerial Scene adaptation tasks. For the proposed tasks, the data for the same scene is collected not only by the traditional camera on the ground, but also by satellite from the out space and unmanned aerial vehicle (UAV) at the high-altitude. Since the semantic gap between the ground/satellite scene and the aerial scene is much larger than that between ground scenes, the newly proposed tasks are more challenging than traditional domain adaptation tasks. The datasets/codes can be found at https://github.com/jianzhelin/DuAN.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**.

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413893

[1] A task-specific classifier means the classifier trained for a specific task such as object classification or semantic segmentation. In this paper, our task is image classification.

KEYWORDS

Domain Adaptation, Ground/Satellite-to-Aerial Scene, Task-specific

ACM Reference Format:

Jianzhe Lin, Lichao Mou, Tianze Yu, Xiaoxiang Zhu, and Z. Jane Wang. 2020. Dual Adversarial Network for Unsupervised Ground/Satellite-to-Aerial Scene Adaptation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413893

1 INTRODUCTION

Recent advances in deep learning not only bring impressive performance for data processing, but also aggravate the burden of data annotation. To train a reliable deep neural network, excessive annotated data with labels are required. This annotation concern is severe for remote sensing data. Nowadays, with much easier access to this type of data, annotation of newly collected remote sensing data has become a big problem, as human labor for annotation is expensive, and limited prior knowledge exists for remote sensing data.

Domain adaptation might solve this problem in a straight forward manner. By domain adaptation, the label-scarce remote sensing data (the target domain) can borrow information directly from the label-rich regular RGB image data (the source domain). As data from such two domains are hard to be aligned, effective adaptation is challenging. This task is even more challenging when the target remote sensing samples are totally unlabeled. In this work, we propose a novel unsupervised domain adaptation (UDA) method to tackle the above challenge.

A popular research direction of UDA is based on adversarial learning, which is to align data with different distributions in an adversarial manner: A feature generator is trained to generate the domain invariant features for both source and target domain samples, in order to fool a domain discriminator which is trained to discriminate the domain labels of the features generated by the generator [1][22].

However, there are two potential limitations of the above adversarial learning based UDA. First, this method might not be task-specific. The adapted target domain data could lose its discriminative data distribution, which is essential for its classification [14][21][11]. The generated aligned feature vectors of the target data might not perform well in task-specific classifiers. Second, the source and target domain data are treated in the same way during the adaptation process. To be more specific, raw data from two different domains pass through a standard feature generator and then a task-specific classifier. Such a process may not be preferred as the data from two domains should serve for different purposes: The target domain data needs to serve for task-specific classifiers, whereas the source domain data should be supplementary. The objective for source domain data is mainly related to the feature adaptation but not to the classification task. To make the two domains function well respectively for their own objective, we proposed the dual adversarial network.

In this work, we assign two domains with domain-specific tasks. The source domain mainly serves for the feature adaptation, whereas the target domain would be task-specific. To achieve the task-specific goal with unlabeled target domain data, we introduce two individual classifiers, which can classify source samples correctly, to provide inconsistent classification results for target domain data simultaneously. The model loss will be generated by the inconsistency to optimize the target domain feature generator. The dual adversarial learning is proposed to complete the domain-specific tasks.

The proposed dual adversarial network(DuAN) includes four players: Two task-specific classifiers, the source feature generator, the target feature generator, and the domain discriminator. A comparison between the proposed DuAN and the classical domain adversarial network can be found in Fig. 2. In the first adversarial learning phase, the source domain feature generator generates features by mimicking the target domain features which are fixed in this phase to fool domain discriminator; For the second adversarial learning phase, task-specific classifiers whose weights are initialized by the source domain features generated in the first phase yield inconsistent classification results to fool the target domain feature generator: let it mistake the two classifiers are for different tasks. Such a feature generator is more like a "task discriminator": It only realizes that the two classifiers are for the same task when the two task-specific classifiers provide the same classification results. These two phases will iterate until the domain discriminator is fooled, and meanwhile the target feature generator does not get fooled. Compared with the traditional adversarial domain adaptation, our source domain feature generator only needs to generate features for the feature adaptation, and thus the generated features are better aligned and adapted; The target domain feature generator, which does not participate in adaptation directly but only plays the adversarial game with classifiers, would generate much more discriminative features. Major contributions of this paper can be summarized as follows:

- We propose separate feature generators to serve for **domain-specific purposes** (e.g., feature adaptation and classification task). The generated target domain features can better preserve the discriminative target domain data distribution.
- We propose the **Dual Adversarial Network (DuAN)**. The network is trained in a stepwise manner. Four "players" play two adversarial games in DuAN, one for the feature adaptation, and the other for the classification task.
- We investigate a novel, challenging **satellite/ground-to-aerial Scene Adaptation task (GSSA)**. This task not only explores the effectiveness of domain adaptation for remote sensing data (satellite-to-aerial), but also aims to solve the label-scarce problem for the aerial scene (ground-to-aerial). Examples of data for GSSA are shown in Fig. 1.

2 RELATED WORK

2.1 Adversarial Domain Adaptation

Recent years have witnessed the exploitation of adversarial domain adaptation, which stems from the technique proposed in [9]. This type of adversarial domain adaptation has one feature generator as well as one domain discriminator [1][26][31]. The generated features from the two domains would be aligned together to fool the domain discriminator until it cannot recognize which domain the features come from. Such aligning in early time was realized by simple batch normalization statistics[17][4][16][5], which aligned the

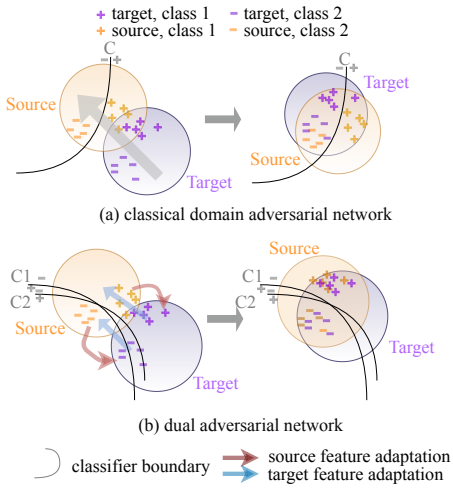


Figure 2: (Best Viewed in color.) Illustration of the mechanism comparison between the classical adaptation approach and the proposed DuAN. (a) The classifier cannot classify target domain data well although two domain data are aligned well, as they might fail to consider task-specific classifiers during adaptation. (b) Two individual task-specific classifiers first trained on the source domain data provide inconsistent classification results for the target domain data. Such discrepancy would be minimized in an iterative way: 1. the source data feature mimics the target data feature, 2. classifiers are updated based on the new source data distribution and provide new discrepancy, 3. the target data feature is updated to minimize such discrepancy. The target data will be suitable for various task-specific classifiers at last.

source and target domain to a canonical one. By further introducing loss to mix up data from both domains, it was more difficult for the domain discriminator to classify the domains [10][25][30][39]. However, such methods were not task-specific, which meant the generated feature might not work well on the classifier [24][2][19]. Recently, the Maximum Classifier Discrepancy (MCD) method was proposed to make the adversarial mechanism to be task-specific by constructing adversarial learning between *task-specific classifiers* and *feature generator* [8][13][12]. To be more specific, two task-specific classifiers at the same time took features from the generator. The feature generator tried to fool the two classifiers by generating ambiguous features for input samples [20], while the two task-specific classifiers would try best to get the uniformed results to avoid being fooled by the feature generator.

However, we have to point out that the MCD framework ignores the effectiveness of the feature generator. The data from the two domains, which are for different tasks (the source domain data is mainly for feature transfer task and the target domain data is mainly for classification task), shouldn't generate features in the same way. The same feature generator for the source and target domain data might not provide powerful uniformed features if the data from two domains are with a large semantic/feature gap. To address this concern in challenging and more practical domain adaptation scenarios, we propose the DuAN method.

2.2 Ground/Satellite-to-Aerial Scene Adaptation (GSSA)

In this work, we mainly want to apply the domain adaptation to remote sensing data. Remote sensing data can be generally divided into *Satellite data* and *Aerial data*. Nowadays, with much easier access to such remote sensing data, its annotation is a highly practical concern. We first explore the relationship between different types of remote sensing data by domain adaptation between the satellite scene and the aerial scene. We then explore domain adaptation to help with the annotation of remote sensing data, taking advantage of the ground scene data. We name these two tasks as GSSA tasks. Examples for such tasks are shown in Fig. 1. We assume that image data captured from different views under the same scene class have consistent underlying intrinsic semantic characteristics, although with a large feature gap. With rich information transferred from the ground view data that can be easily obtained from ImageNet [6] or SUN [35], the understanding and annotation of label-scarce aerial data can be better served.

Previously, works for addressing this cross-view (ground-to-aerial) domain adaption problem was mainly based on image geolocalization [33]. There were also works [28][29][27][7] that assumed the scene transfer from ground to aerial as a particular case of cross-domain adaptation, in which the divergences across domains were caused by viewpoint changes. However, all existing methods were based on relatively simple models and were tested on small datasets. There is no existing benchmark for this challenging task. In this paper, we for the first time propose a uniformed GSSA benchmark for the domain adaptation task.

3 MODEL

In this section, an overview of the proposed Dual Adversarial Network (DuAN) is given to present a comprehensive picture. Afterward, the model initialization and training are described respectively.

3.1 Overview

As illustrated in Fig. 3. Five components exist in our framework: the domain discriminator D_1, the source feature generator G_1, the target feature generator G_2, the classifier C_1, and the classifier C_2. The general process is separated into two parts, model initialization and parameter learning. The feature generators G_1 and G_2, and the domain discriminator D_1 are initialized by adversarial learning, while classifier C_1, C_2 are initialized by classification on the source domain features. Parameters of every component are learned in a stepwise manner. First, G_2 as "task discriminator" is optimized based on classification discrepancy between C_1 and C_2, and the output feature of G_2 is updated; Second, the parameters of G_1 and D_1 are optimized by feature discrepancy between the newly generated G_2 feature and the former G_1 feature, and the new G_1 feature generated; Third, C_1 and C_2 are optimized by the cross-entropy loss based on the G_1 feature. The updated C_1 and C_2 will further return to step one to update G_2. The three steps will iterate until convergence. In the process, G_2 is fully task-specific, whereas the major task of G_1

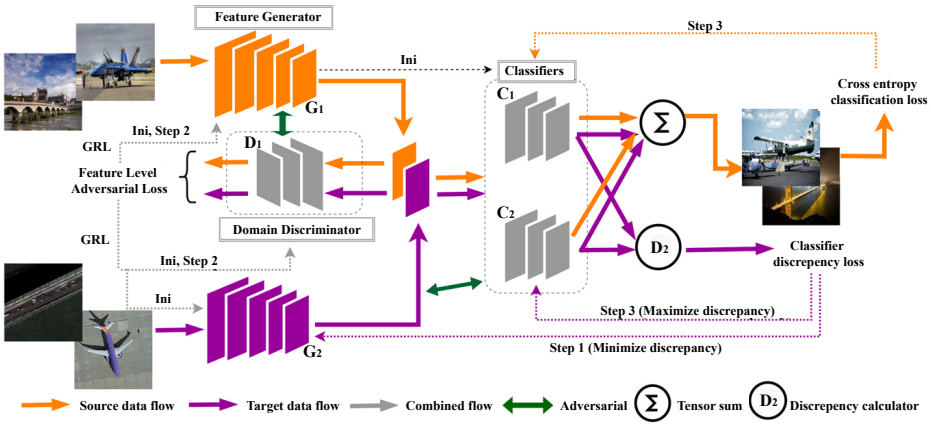


Figure 3: The flowchart of the proposed DuAN. Two adversarial processes exist, where one for the feature adaptation is realized by the source flow (orange color), and the other for the classification task is realized by the target flow (purple color). Flow here means the forward and backward propagation in the neural network. Steps 1-3 refer to the three iterative training steps. Components in the corresponding step are updated iteratively. "Ini" is the abbreviation for model initialization.

is to generate features of the source domain to mimic target domain features. These three steps are illustrated in Fig.3.

The inputs of general framework is formulated as follows. The labeled source domain data is represented with $X_s = \{x_s^i, y_s^i\}_{i=0}^{N_s}$, and the unlabeled target domain data is represented with $X_t = \{x_t^i\}_{i=0}^{N_t}$. N_s and N_t represent the numbers of data on the two domains respectively. The source domain feature set $F_s = \{f_s^i, y_s^i\}_{i=0}^{N_s}$ with known labels y_s is first generated by $f_s = G_1\{x_s; \theta_{G_1}\}$, in which θ_{G_1} means the parameters of G_1. The target domain feature set is generated by $f_t = G_2\{x_t; \theta_{G_2}\}$ in which G_2 is the target feature generator and θ_{G_2} means its parameters.

3.2 Model Initialization

The model is first initialized conventionally. The source and target domain features are the inputs to the domain discriminator, which is represented as $D_1\{f_s, f_t; \theta_{D_1}\}$. The two generators would try to fool D_1 while D_1 would be maximized to classify the features' domain labels. At the same time, the two classifiers assign labels to the source domain features, based on the regular cross-entropy loss. These two classifiers are formulated as $C_1\{f_s; \theta_{C_1}\}$ and $C_2\{f_s; \theta_{C_2}\}$. Our first min-max objective is

$$\min_{\theta_{C_1},\theta_{C_2}} \max_{\theta_{G_1},\theta_{G_2},\theta_{D_1}} \alpha_1 \mathcal{L}_{d_1}(D_1, G_1, G_2) + \beta_1 \mathcal{L}_{t_1}(G_1, C_1, C_2), \tag{1}$$

where α_1 and β_1 are weights for the two losses, and we also define $\mathcal{L}_{d_1}$ and $\mathcal{L}_{t_1}$ as

$$\mathcal{L}_{d_1}(D_1, G_1, G_2) = \mathrm{E}_{x^t}\left[\log D_1(G_2(\mathrm{x}^t;\theta_{G_2});\theta_{D_1})\right] + E_{f^s}\left[\log(1 - D_1(f^s;\theta_{D_1}))\right], \tag{2}$$

$$\mathcal{L}_{t_1}(C_1,C_2,G_1) = \mathrm{E}_{f^s,y^s,z}\left[-y^{s\mathrm{T}} \log C_1(f^s;\theta_{C_1})\right] + \mathrm{E}_{f^s,y^s,z}\left[-y^{s\mathrm{T}} \log C_2(f^s;\theta_{C_2})\right], \tag{3}$$

where y^s means the one-hot encoding of the labels of source domain data. In both equations, $f^s = G_1(x^s, z; \theta_{G_1})$ as defined earlier. In our implementation, for both G_1 and G_2, we use resnet to extract the features, and D_1, C_1, and C_2 are regular resnet classifiers. For the above minmax objective, we solve the problem by updating $\theta_{G_1}, \theta_{G_2}$ (freezing $\theta_{D_1}, \theta_{C_1}, \theta_{C_2}$) and $\theta_{D_1}, \theta_{C_1}, \theta_{C_2}$ (freezing $\theta_{G_1}, \theta_{G_2}$) alternatively. We can initialize all parameters of the proposed model in this way.

3.3 Model Training

After the initialization of the model parameters, we can get differed classification results from C_1 and C_2. The following model training is divided into three steps.

Step 1 and classifier discrepancy loss: In this step, we use the discrepancy loss to train the target feature generator G_2, while other components are frozen. The two classifiers try to fool G_2 with inconsistent classification results whereas G_2 tries to generate the features to make them look the same to avoid being fooled. Here we introduce D_2 to identify the difference between the results of two classifiers. D2 is only an identifier with no parameters. The objective of this step is to minimize the discrepancy loss defined in Eq. 4 as

$$L_{d_2}(D_2, C_1, C_2) = D_2(C_1(\mathrm{f}^t;\theta_{C_1}), C_2(\mathrm{f}^t;\theta_{C_2})). \tag{4}$$

Here $\mathcal{L}_{d_2}$ is the discrepancy loss between the two classifiers. The only variable in this step is θ_{G_2}. For D_2, different from D_1 which is defined by the neural network, it is just an identifier which is defined as

$$D_2(x, y) = \frac{1}{n}\sum_{n=1}^{N} |\mathrm{x}_n - y_n|, \tag{5}$$

in which N is the total number of elements for x and y (x and y should have the same number of elements). We use the L-1 norm to calculate the difference between the two inputs.

Step 2 and feature adversarial loss: In this step, we train the feature generator G_1 and the domain discriminator D_1 in an adversarial manner, with all other components being frozen. Different from the traditional UDA, only features from the feature generator G_1 are updated to appear as if generated from G_2, to fool D_1 which would try best to discriminate the features from two domains. The objective of this step is to minimize the discrepancy between source and target domain features by D_1, which is formulated in Eq. 6 as

$$\min_{\theta_{G_1},\theta_{D_1}} \mathcal{L}_{d_1}(D_1, G_1, G_2), \tag{6}$$

in which $\mathcal{L}_{d_1}$ is the feature adversarial loss defined in Eq. 2. Such loss will optimize the network parameters in a Gradient Reverse Learning (GRL) [9] way, as higher loss means worse adaptation performance. The variables to be optimized in this step are θ_{G_1} and θ_{D_1}. After this step, the feature output of G_1 is updated, which will be used to optimize the classifiers. However, as G_2 is not involved in this step, its generated target domain feature is only related to the classification task.

Step 3 and cross-entropy loss: In this step, we train C_1 and C_2 with other components being frozen. This step has two objectives, the first is to make the two classifiers as dissimilar as possible, for the adversarial purpose as in Step 1. The second objective is to maximize the classification accuracy of both classifiers for features from G_1 by minimizing cross-entropy losses, which is a task-specific objective. To jointly consider these two objectives, the objective function is defined as

$$\begin{aligned}\max_{\theta_{C_1},\theta_{C_2}} \; & \alpha_2 L_{d_2}(D_2, C_1, C_2) \\ & + \beta_2 L_{t_2}(G_1, C_1, C_2),\end{aligned} \tag{7}$$

where α_2 and β_2 are weights for the two losses, and $\alpha_2/\beta_2 = \alpha_1/\beta_1$. We define $\mathcal{L}_{t_2}$ the same as $\mathcal{L}_{t_1}$ in Eq.3, and define $\mathcal{L}_{d_2}$ the same as in Eq. 4.

For both C_1 and C_2, the input are the features from G_1 and G_2.

Dual Adversarial Network Training: Detailed training process for the Dual Adversarial Network can also be found in Alg. 1. In the algorithm, two adversarial parts exist. The first is between step 1 and step 3, and the other is in step 2 between G_1 and D_1. The three steps would iterate, not only until the classification results on G_1 are converged but also until: 1. The D1 gets fooled by G_1 and cannot discriminate which domain the data are from; 2. The G_2 does not get fooled by C_1 and C_2, and realizes that the two classifiers are for the same task. We want to point out that these three steps cannot be integrated into one step, as step 1 and step 3 are with adversarial objectives, and inputs of the three steps are different. However, the order of the three steps is not important. These three steps will iterate until convergence.

4 EXPERIMENTS

In the experimental part, we conduct our experiments on three tasks. The first is the Visual Domain Adaptation Challenge (VisDA) 2017 challenge for image classification, the second is domain adaptation between two types of remote sensing scene (namely the satellite scene and the aerial scene), in order to explore the relationship between them. The third is the Ground-to-Aerial scene Adaptation task, which is the most challenging. Below we will first describe the datasets.

Algorithm 1: Training for DuAN.

Input: image normalization for both the source and the target domain data;
Output: the optimized weights for G_1, G_2, D_1, C_1, C_2;
while *epoch* $\leq$ *max epoch* **do**
  for $batch \leftarrow 1$ **to** N **do**
    Step 1: Input the normalized target domain data with index $(N+1)/2 \rightarrow N$, optimize G_2 by minimizing Eq. 4;
    Step 2: Input the normalized source domain data with index $0 \rightarrow N/2$, optimize G_1 and D_1 by minimizing Eq. 6;
    Step 3: Input the normalized source domain data with index $0 \rightarrow N/2$, optimize C_1 and C_2 by maximizing Eq. 7.
  end
end

4.1 Datasets and Setup

VisDA 2017 challenge: we first evaluate our proposed DuAN model on the VisDA 2017 challenge. A detailed introduction can be found in the supplementary material.

Satellite to aerial scene adaptation: For this task, we collect nine classes for domain adaptation, including River, Parking lot, Overpass, Harbor, Forest, Building, Beach, Residential, Agricultural. The datasets are mainly collected from the WHU-RS dataset [34], the UCMerced dataset [38], as well as the data collected by ourselves online and through our collaborators. The data from the satellite view is with much lower resolution and clarity when compared with the data from the aerial view. The data are re-scaled to the resolution of 256×256. There are 53 images/class for the source domain, and 100 images/class for the target domain, and in total 1377 images. A visualized comparison of these two types of remote sensing data is shown in the left of Fig. 4.

Ground to aerial scene adaptation: For this task, we include 15 classes, as shown in Fig. 4. Each image is re-scaled to the resolution of 256×256. Each class has 5,800 images (5,000 from the source domain and 800 from the target domain), and the datasets contain 87,000 images in total. We randomly choose 25,000 images from the source domain for training, and use the trained model to test on the validation data, which are randomly chosen 5% target domain data. After validation, we use the rest target domain data for testing. For this task, the data from the ground view has a huge distribution gap when compared with the data from the aerial view, as can be noted in the examples. This task is highly challenging. Moreover, the similarity between classes in the same view also makes this task difficult. For example, the features of the parking lot are similar to that of the harbor, and the runway looks similar to the bridge from the aerial view. Also, we need to point out that "water park" corresponds to "water plant". We use these similar classes to set up the pairs for this class due to the lack of "water park" in the aerial scene. Data examples for this task are shown in the right of Fig. 4.

For the network setup, we used Adam optimizer with learning rate 2×10^{-4} with no decay. The batch size is set to 64. For Eq. 1 and Eq. 7, $\alpha/\beta = 0.1$. The experiment results comparison of different α/β can be found in Sec. 4.5. The above parameter settings are suitable for all scenarios.

4.2 VisDA Challenge

4.2.1 Setup. In this experiment, we first evaluated the performance of the proposed DuAN model on the VisDA 2017 challenge. Due to a large number of image data for training, we run the experiments on our server. For the hardware, the CPU is AMD Ryzen 2nd

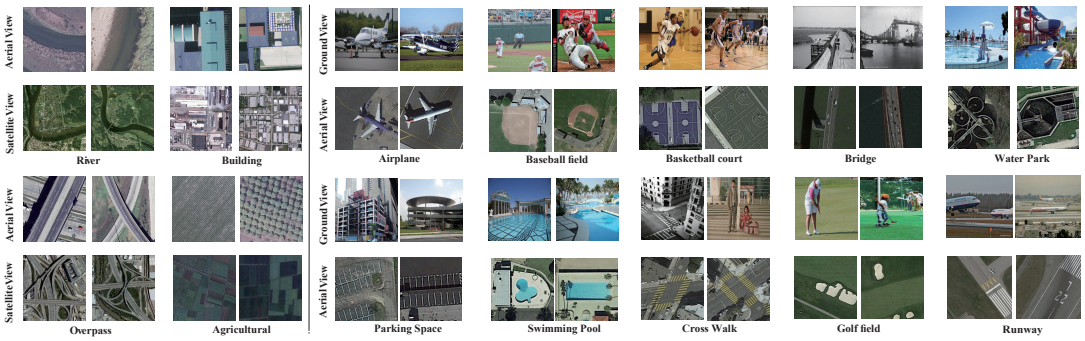
Table 1: Accuracy(%) results for the VisDA 2017 Challenge task with ResNet-101 as the base network

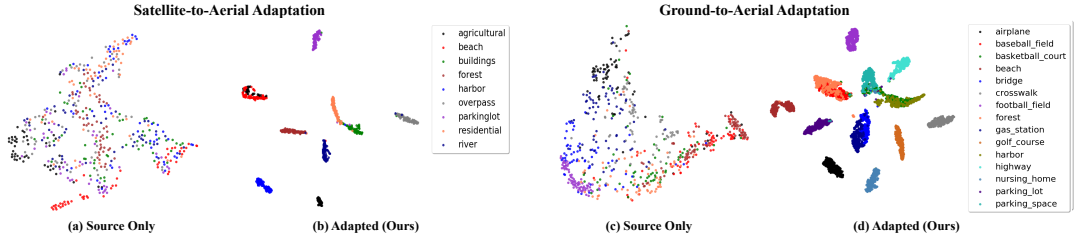
Method	Plane	Bcycl	Bus	Car	Horse	Knife	Mcycl	Person	Plant	Sktbrd	Train	Truck	Average
Source	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN [18]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	*85.8*	20.7	61.1
DANN [9]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [8]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
HAFN [36]	92.7	55.4	82.4	70.9	93.2	71.2	90.8	78.2	89.1	50.2	88.9	24.5	73.9
SAFN [36]	93.6	61.3	**84.1**	70.6	94.1	79.0	91.8	**79.7**	9.9	55.6	**89.0**	24.4	76.1
DuAN	**96.4**	**84.3**	80.9	**82.4**	**97.3**	**86.9**	**92.1**	77.4	**92.5**	**78.4**	74.1	**29.2**	**80.2**

Table 2: Accuracy(%) results for the Satellite-to-aerial Scene Adaptation task with ResNet-101 as the base network

Method	River	Parking lot	Overpass	Harbor	Forest	Building	Beach	Residential	Agricultural	Average
Source	53.0	0.0	0.0	4.0	44.0	14.0	0.0	22.0	52.0	21.0
DANN [9]	53.0	0.0	0.0	92.0	98.0	0.0	0.0	0.0	28.0	24.2
PADA [3]	75.0	94.0	83.0	84.0	50.0	21.0	83.0	80.0	69.0	71.0
MEDA [32]	**93.0**	96.0	64.0	96.0	78.0	51.0	93.0	88.0	83.0	82.4
JADA [15]	91.0	93.0	63.0	96.0	54.0	67.0	95.0	83.0	76.0	79.7
HAFN [37]	75.0	91.0	58.0	90.0	79.0	33.0	86.0	70.0	73.0	72.7
SAFN [36]	64.0	**100.0**	67.0	95.0	**100.0**	70.0	99.0	60.0	**94.0**	83.2
MCD [8]	84.0	**100.0**	65.0	**100.0**	**100.0**	51.0	**100.00**	79.0	89.0	85.3
SWD [13]	90.0	**100.0**	53.0	92.0	59.0	23.0	96.0	80.0	74.0	74.1
DTA [14]	87.0	76.0	89.0	91.0	91.0	62.0	96.0	76.0	78.0	82.9
DuAN	49.0	98.0	**91.0**	**100.0**	99.0	**100.0**	99.0	**90.0**	**94.0**	**91.1**

Figure 4: Left: Examples from the proposed Satellite-to-aerial domain adaptation datasets with 9 categories. Right: Examples from the proposed ground-to-aerial domain adaptation datasets with 15 categories (except for classes in Fig. 1).





(a) Source Only (b) Adapted (Ours) (c) Source Only (d) Adapted (Ours)

Figure 5: (a)-(b): t-SNE [23] visualization results of domain adaptation methods for the Satellite-to-aerial scene adaptation. (c)-(d): t-SNE [23] visualization results of domain adaptation methods for the Ground-to-aerial scene adaptation. We can see that after applying our adaptation methods, the target samples are more discriminative.

Threadripper 2990WX, and GPU is NVIDIA RTX TITAN × 2, with 128GB Memory. This hardware also works for the ground-to-aerial scene adaptation task. We select ResNet-101 for this task as the base network. All comparison methods are trained until convergence.

4.2.2 Results. Table 1 reports our results and also the results obtained from previous studies. We directly compare our results with the reported results in previous papers to make the comparison fair. As this part is only for method verification, we only make comparisons with the representative methods DAN [18] and DANN [9], our baseline method MCD [8], and the most recently proposed method HAFN [36] and SAFN []. We can find from the table that DuAN is always with the best performance from the perspective of average accuracy, followed by SAFN, HAFN, MCD, and others. DANN and DAN also get excellent performance for specific classes like Bus and Train.

4.3 Satellite-to-aerial Scene Adaptation

4.3.1 Setup. We run the experiments locally on computer, as the dataset for satellite-to-aerial scene adaptation is not large. For the hardware, the CPU we adapt is Intel® Core^{TM} i7-8700k, and GPU we use is NVIDIA GEFORCE GTX 1080 TI. For this task, we adapt ResNet-101 as our basenet. We implement each comparison method all by ourselves, including the first adversarial domain adaptation work DANN [9], the recent SOTA PADA [3], MEDA [32], JADA [15], HAFN [37], and SAFN [37] based on DANN, as well as three SOTA task-specific methods [8], DTA [14], and [13]. The work in [13] and DTA [14] are generally modifications of MCD [8]. Therefore, we choose MCD as our major comparison method. We provide not only detailed accuracy comparison for each method but also a visualized t-sne comparison for the target data before (source only) and after adaptation by our method, as shown in Fig. 5. All methods have been trained for 100 epochs, as testing accuracy of every method has converged at such epoch number. The trained model for every ten epochs is tested directly on target domain data without validation as the size of the dataset is not big, and the best performance is reported to make a comparison.

4.3.2 Results. As can be found in Table 2, the proposed method DuAN is with the best overall accuracy, followed by MCD, and SAFN. The accuracies of DANN and source only are both around 30%. Also, we can find the building class is most difficult for classification as it is quite easy to be confused with Residential class. But for our method, the accuracy of this class is 100%. By accuracy comparison with source only method, we can find the two types of remote sensing scenes can be aligned by domain adaptation, which proves that the information between different types of remote sensing images can be shared and exchanged. Also, it can be concluded from t-SNE comparison in Fig. 5: Although the target samples do not separate well in the non-adapted situation, they do separate clearly in the adapted situation. Such a conclusion proves the significance of the proposed satellite-to-aerial adaptation task, as information transfer between these two types of images can help with their classification.

4.4 Ground-to-aerial Scene Adaptation

4.4.1 Setup. For this task, we use ResNet-101 as the base network. We show the detailed accuracy comparison in Table 3. The training epochs are always set as 30, as all settings would converge at this epoch number. For all comparison methods and the proposed method, as the target domain dataset is large, we use 5% randomly selected target domain data for validation and the rest for testing.

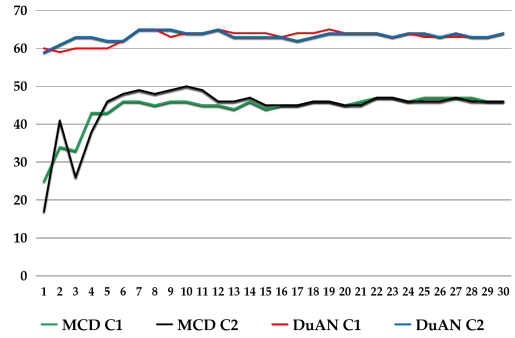


Figure 6: The classification accuracies on validation data.

The trained model parameters with the lowest loss on the validation phase are used for testing. We also make detailed visualized t-sne results comparison as in Fig. 5.

4.4.2 Results. As noted in Table 3, the overall accuracy (OA) of the proposed method is 64.40%, much higher than that of other methods which are mostly lower than 50%. In the table, basketball court, baseball field, water park, parking lot and parking space are abbreviated as basketball., baseball., water., parking.L and parking.S respectively. We want to provide two observations for this result. First, there is an indoor class baseball field for the ground scene but outdoor for the aerial scene. Therefore, this class is with larger domain gaps than the other classes. All methods fail and get the lowest classification accuracy for this class. Second, there is a possibility that the source domain data are more discriminative than target domain data. Two representative classes are the swimming pool and the basketball field, for which aerial view data are easily be mistaken and classified as the water park and the golf field respectively. For these classes, the loss of discriminative distribution of target domain data during the adaptation process might even result in better performance. Such observation can explain our failure in these classes compared with other methods. Also, this observation can tell the low accuracies come from false classification instead of random data noise. The t-SNE comparison between the adapted result and the source-only result proves the effectiveness of domain adaptation.

4.4.3 Model Training Observations. We take the ground-to-aerial adaptation task as an example to demonstrate the advantage of our proposed method in terms of model training. Fig.6 shows the changes in classification accuracy on validation data at different epochs. For this task, the training times based on our hardware settings as mentioned in Sec. 4.2.1 is 10.5 mins/epoch. We also use our baseline method MCD [8] for comparison, whose training time is 10.1 mins/epoch. We use the model parameters trained at the epochs with the highest accuracies on validation data to do the testing for DuAN and MCD. We would like to mention two observations. First, **from the perspective of convergence of classification**, due to our stepwise model training, the classification result of the proposed DuAN stops to change at the 5th epoch while the result of MCD takes much longer to get converged. Also, at the first epoch, DuAN already yields a quite high accuracy. We need to point out

Table 3: Accuracy(%) results for the Ground-to-aerial Scene Adaptation task with ResNet-101 as the base network

Method	Airplane	Baseball.	Basketball.	Beach	Bridge	Crosswalk	Forest	Golf	Harbor	Parking.L	Parking.S	Residential	Runway	Swimming	Water.	Average
Source	0.25	0.38	6.62	0.00	27.25	49.88	70.88	0.00	1.50	0.25	0.12	0.00	0.00	17.38	2.00	11.77
DANN [9]	35.38	**1.00**	**17.50**	0.00	0.25	49.25	0.00	5.62	0.00	1.00	0.25	0.12	0.50	41.38	0.12	10.16
PADA [3]	39.43	0.26	25.03	66.89	52.46	43.24	21.58	46.37	3.43	28.34	21.57	13.44	2.94	4.66	1.25	24.73
MEDA [32]	39.63	0.13	5.87	82.72	70.82	3.85	96.13	72.24	43.31	28.51	36.46	64.04	32.53	10.30	1.21	39.18
JADA [15]	90.15	0.33	8.32	91.63	**96.82**	26.53	96.37	92.47	41.39	65.44	33.56	61.04	31.53	10.30	2.57	49.90
HAFN [37]	89.36	0.47	2.03	63.45	87.82	5.68	94.67	89.46	61.48	62.73	40.51	54.01	20.33	4.34	0.62	45.13
SAFN [37]	92.14	0.85	6.83	68.70	45.41	56.56	81.33	44.44	82.60	79.40	42.56	91.34	35.43	**75.50**	**2.44**	54.70
MCD [8]	71.38	0.38	0.38	**100.0**	91.38	0.00	**100.0**	**99.62**	0.75	45.12	44.50	83.50	40.62	71.27	1.75	45.73
SWD [13]	50.04	0.27	5.03	80.72	77.82	0.00	94.67	82.12	3.31	29.53	46.46	61.04	40.53	15.30	1.57	39.09
DTA [14]	87.11	0.34	3.63	81.42	69.49	51.08	96.39	72.44	35.43	79.44	49.56	86.04	41.42	8.41	0.91	50.87
DuAN	**99.38**	0.25	4.62	**100.0**	2.50	**100.0**	**100.0**	97.88	**99.88**	**100.00**	**75.00**	**96.50**	**89.38**	0.12	0.62	**64.40**

that for almost all UDA methods, the accuracy results are highest at the first or second epoch and then reduce a bit. Second, **from the perspective of convergence of adaptation**, the discrepancy between C1 and C2 in DuAN decreases much faster than in MCD. As in DuAN, each domain is assigned a specific task, the classifiers can get consistent results much faster than in MCD. This suggests that the adaptation process can get converged much faster in a stepwise manner, and we can obtain uniformed task-specific classification results in much shorter time.

4.5 Ablation Study

In the ablation study part, we not only verify the effectiveness of each component of our network, but also compare different parameters settings of α/β (α_1/β_2 in Eq. 1, and α_2/β_2 in Eq. 7). We choose both the task Satellite-to-Aerial scene adaptation (StA-DA) and the task Ground-to-Aerial scene adaptation (GtA-DA) to complete the experiment. We use the average classification accuracy as the criterion for evaluation.

For the ablation study of each network component, the experimental result can be found in Table 4. From the model level, we separate our model to three parts, domain-specific feature generator(D-SFG, we compare with one common feature generator for ablation study of this component), domain discriminator(DD), and two classifiers (TC, we compare with one classifier for ablation study of this component). From the loss level, our network mainly includes the the feature adversarial loss L_{d_1}, classifier discrepancy loss L_{d_2}, and the cross-entropy loss L_t ($L_t = L_{t_1} = L_{t_2}$ as described in Sec. 3.3). We need to notice that DD and L_{d_1} cannot be separated. The same goes for TC and L_{d_2}.

We can find from the table, the first setting can be regarded as the regular domain adversarial adaptation with a common feature generator, a domain discriminator, and a classifier. The second setting is the same as the MCD method, and for the third we add a domain-specific feature generator. The last is the setting for the proposed method.

For the selection of parameter α/β, a comparison can be found in Table 5. We select the representative numbers for α/β to make the comparison.

Table 4: Mean value comparison.

Model			Loss			StA-DA	GtA-DA
D-SFG	DD	TC	L_t	L_{d_1}	L_{d_2}	Accuracy	Accuracy
×	✓	×	✓	✓	×	24.22	10.16
×	×	✓	✓	×	✓	85.33	45.73
✓	×	✓	✓	×	✓	82.00	53.16
✓	✓	✓	✓	✓	✓	**91.11**	**64.40**

We have to point out that we might have other choices of numbers for the α/β (e.g. 0.5), but the performance does not change too much. Therefore, based on the results in Table 5, we set α/β as 0.1.

Table 5: The classification results comparison for different α/β.

α/β	0.001	0.1	1	10	100
StA-DA	46.00	**91.11**	86.44	82.70	77.22
GtA-DA	23.54	**64.40**	55.63	54.21	46.37

5 CONCLUSION

In this paper, we propose a novel adversarial domain adaptation model, named Dual Adversarial Network (DuAN), motivated by the idea that the source and target domain data should not be treated in the same way in domain adaptation. Different from previous methods, we propose a domain-specific strategy for the feature adaptation and the classification task, in order to relieve the loss of discriminative characteristics of the target domain data during the adaptation process. The model is optimized in a stepwise manner. We also propose a novel "Ground/Satellite-to-Aerial Scene Adaptation" task. This adaptation task is for a highly challenging and practical scenario with larger domain gap when compared with traditional domain adaptation tasks. Also, such an adaptation can help to tackle the remote sensing data automatic annotation problem. The superior experiment results for both VisDA 2017 challenge and GSSA task prove the effectiveness of our proposed method.

REFERENCES

[1] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. 2016. Domain Separation Networks. In *NIPS*. 343–351.

[2] Z. Cao, M. Long, J. Wang, and M. I. Jordan. 2018. Partial Transfer Learning with Selective Adversarial Networks. In *CVPR*.

[3] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. 2018. Partial adversarial domain adaptation. In *ECCV*. 135–150.

[4] F. Maria Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. Rota Bulo. 2017. Autodial: Automatic domain alignment layers. In *CVPR*. 5067–5075.

[5] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. 2019. Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. In *CVPR*. 7354–7362.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. 2009. Imagenet: A large-scale hierarchical image database. , 248–255 pages.

[7] Zhipeng Deng, Hao Sun, and Shilin Zhou. 2018. Semi-Supervised Ground-to-Aerial Adaptation with Heterogeneous Features Learning for Scene Classification. *ISPRS International Journal of Geo-Information* 7, 5 (2018), 182.

[8] Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. 2018. K. Saito and K. Watanabe and Y. Ushiku and T. Harada. In *CVPR*. 3723–3732.

[9] Y. Ganin and V. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. , 1180–1189 pages.

[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation.

[11] Vinod Kumar Kurmi, Shanu Kumar, and Vinay P. Namboodiri. 2019. Attending to Discriminative Certainty for Domain Adaptation. In *CVPR*. 491–500.

[12] Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. 2019. Unsupervised domain adaptation based on source-guided discrepancy. In *AAAI*. 4122–4129.

[13] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*. 10285–10295.

[14] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. 2019. Drop to Adapt: Learning Discriminative Features for Unsupervised Domain Adaptation. In *ICCV*. 91–100.

[15] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. 2018. Joint Adversarial Domain Adaptation. In *ACM MM*. 729–737.

[16] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. Adaptive batch normalization for practical domain adaptation. *PR* 80 (2018), 109–117.

[17] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. 2016. Revisiting batch normalization for practical domain adaptation. In *arXiv:1603.04779*.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015).

[19] M. Long, Z. Cao, J. Wang, and M. I.Jordan. 2018. Conditional Adversarial Domain Adaptation. In *NIPS*. 1640–1650.

[20] Ping Luo, Fuzhen Zhuang, Hui Xiong, Yuhong Xiong, and Qing He. 2008. Transfer learning from multiple source domains via consensus regularization. In *CIKM*. 103–112.

[21] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation. In *CVPR*. 2507–2516.

[22] Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. 2019. GCAN: Graph Convolutional Adversarial Network for Unsupervised Domain Adaptation. In *CVPR*. 8266–8276.

[23] L.v.d Maaten and G. Hinton. 2008. Visualizing data using t-sne. *JMLR* 9, 6 (2008), 2579–2605.

[24] Z. Pei, Z. Cao, M. Long, J. Wang, and J. Wang. 2018. Multi-Adversarial Domain Adaptation. In *AAAI*.

[25] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*. 8503–8512.

[26] B. Sun, J. Feng, and K. Saenko. 2016. Return of Frustratingly Easy Domain Adaptation.

[27] Hao Sun, Zhipeng Deng, Shuai Liu, and Shilin Zhou. 2016. Transferring ground level image annotations to aerial and satellite scenes by discriminative subspace alignment. In *IGARSS*. 2292–2295.

[28] Hao Sun, Shuai Liu, Shilin Zhou, and Huanxin Zou. 2015. Transfer sparse subspace analysis for unsupervised cross-view scene model adaptation. *IEEE JSTARS* 9, 7 (2015), 2901–2909.

[29] Hao Sun, Shuai Liu, Shilin Zhou, and Huanxin Zou. 2015. Unsupervised cross-view semantic transfer for remote sensing image classification. *IEEE GRSL* 13, 1 (2015), 13–17.

[30] Youssef Tamaazousti, Hervé Le Borgne, Céline Hudelot, Mohamed El Amine Seddik, and Mohamed Tamaazousti. 2019. Learning more universal representations for transfer-learning. *IEEE T-PAMI* (2019).

[31] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. 2017. Adversarial discriminative domain adaptation. In *CVPR*. 7167—-7176.

[32] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. 2018. Visual domain adaptation with manifold embedded distribution alignment. In *ACM MM*. 402–410.

[33] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In *ICCV*. 3961–3969.

[34] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. 2010. Structural high-resolution satellite image indexing. In *ISPRS*.

[35] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*. 3485–3492.

[36] R. Xu, G. Li, J. Yang, and L. Lin. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. *In The 2019 IEEE International Conference on Computer Vision (ICCV)* (2019).

[37] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1426–1435.

[38] Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 270–279.

[39] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. 2019. Aligning Domain-Specific Distribution and Classifier for Cross-Domain Classification from Multiple Sources. In *AAAI*. 5989–5996.